2005 PASS Community Summit
Microsoft SQL Server Users Conference & Expo

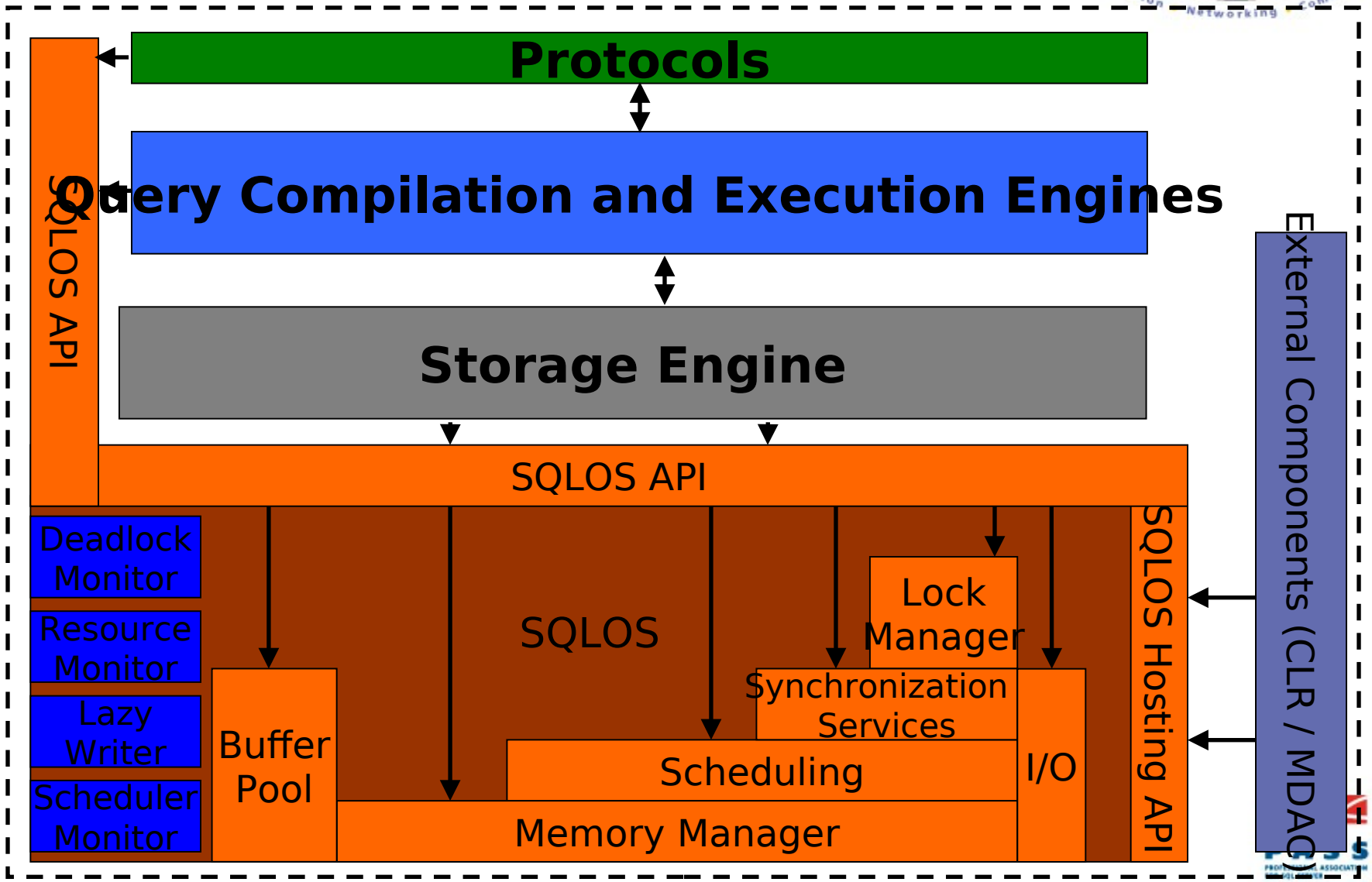
SQLOS

Robert Dorr

SQL Server 2005 OS Foundational Elements

Core SQLOS Ownership

- Scheduling
- Memory
- Buffer Pool
- Hosting
- Exception Handling and User Dumps
- Lock Monitor
- Locking Primitives
- Latching Primitives


Protocols
Query Compilation and Execution Engines
Storage Engine
SQLOS API
SQLOS API
External Components (CLR / MDAC)
Deadlock Monitor
Resource Monitor
Lazy Writer
Scheduler Monitor
SQLOS
Buffer Pool
Lock Manager
Synchronization Services
Scheduling
I/O
Memory Manager
SQLOS Hosting API
= thread

Dedicated Admin Connection (DAC)

- Only SQL Administrator allowed to connect
- Separate Listen On
- TCP Only
- Local only by default
- On clusters turn on ‘remote admin connections’
- Use sqlcmd.exe paramter –A to connect
- Not for everyday use
- Separate reserved memory area

DAC Do's and Don'ts

- Use simple DMV queries
- Avoid memory intensive queries
- Avoid DDL & DBCC Check commands
- Avoid using old SQL 2000 DMV's
- Some functionality is disabled

DAC and DMV History

- **dbcc sqlperf(umsstats)**
- **Debugger poi(poi()) type of activities**
  - Debugger extensions
  - Patent (20040210872) for DSCRIPT technology
    - `… openquery(.. 'sysprocesses.js; "c:\temp\SQLDump0001.mdmp"' …)`
  - SQLDumper.exe
  - Filtered Dumps
    - Approach accuracy of full memory dump
    - Approach speed of mini-dump
    - Size is commonly 100MB or less
    - Faster collaboration with Microsoft
  - The "Transparent Server"
- Message 17883 Introduction

SQLOS DMVs
"Use With Care"

Use of DMVs and DMFs for SQLOS should be limited as needed only.

The facilities have been tested well but the information provided requires thread safe access to critical lists and repeated access to certain lists could have performance implications.

Global Visibility: DMVs

- sys.dm_os_sys_info
- sys.dm_os_loaded_modules
- sys.dm_os_threads
- sys.dm_os_ring_buffers
- sys.dm_os_hosts
- sys.dm_os_stacks

sys.dm_os_sys_info

- Contains
  - Current tick counts
    - cpu_ticks
    - ms_ticks = *GetTickCount*
  - Physical and Virtual Memory Bytes
  - Committed Buffer Pool
  - Max Workers
  - Scheduler Count
  - More…

sys.dm_os_loaded_modules

- Uses EnumProcessModules API (psapi.h)
- Contains
  - Name
  - Versions
  - Company
  - Loaded Address
  - More…
- Ex: lm / lmv in debugger
- Ex: TLIST detailed output
- WARNING: Loader Lock Access Required

sys.dm_os_threads

- DLLMain tracks every thread creation (SOSBOOT.dll)
- Contains
  - Thread Base Address
  - Creation Time
  - User and Kernel Mode Times
  - Handle
  - Loader Lock Access
  - Priority
  - Affinity
  - SQL Worker
    - Worker Address
    - Scheduler Address
- !runaway

sys.dm_os_ring_buffers

- **Several Types**
  - RING_BUFFER_EXCEPTION
    (Ex: Exception Event in Profiler)
  - RING_BUFFER_SCHEDULER
    (Ex: Actual context switch order of workers)
  - RING_BUFFER_RESOURCE_MONITOR
    (Activity of Resource Monitor)
  - RING_BUFFER_OOM
    (Out of Memory Activity)
  - RING_BUFFER_SCHEDULER_MONITOR
    (Activity of Scheduler Monitor)
  - RING_BUFFER_BUFFER_POOL
    (Activity of Buffer Pool)

- **XML Details About Each Type**

RING_BUFFER_EXCEPTION

```
<Record id = "156" type
  ="RING_BUFFER_EXCEPTION" time
  ="1029341612"><Exception><Task
  address="0x025FD018"></Task><Error>208</Error
  ><Severity>25</Severity><State>7</
  State><UserDefined>0</UserDefined></
  Exception><Stack><frame id =
  "0">0X004B7EAE</frame>……
```

- Examples
  - Time: Maps to sys.dm_os_sys_info
  - Error, Severity and State: Actual error code information
  - Stack: Maps to sys_dm_os_stacks

RING_BUFFER_SCHEDULER

```
<Record id = "1852794" type
  ="RING_BUFFER_SCHEDULER" time
  ="1061279474"><Scheduler
  address="0x003FC040"><Action>SCHEDULER_SWITCH
  _CONTEXT</Action><CPUTicks>2751897676551270</
  CPUTicks><TickCount>1061090953</
  TickCount><SourceWorker>0x03D160E8</
  SourceWorker><TargetWorker>0x03D160E8</
  TargetWorker><WorkerSignalTime>1061090953</
  WorkerSignalTime><DiskIOCompleted>0</
  DiskIOCompleted><TimersExpired>1</
  TimersExpired></Scheduler></Record>
```

sys.dm_os_hosts

- Hosted components
  - MDAC
  - CLR
  - More…
- Contains
  - Type and Name
  - Tasks
  - IO Requests
  - More…

sys.dm_os_stacks

- Extended diagnostics can use this to track actual stacks
  - Ex: Memory Allocations
- Hashed to reduce footprint

```
stack_address      frame_index frame_address
------------------ ----------- -------------
0x0270E1D0         0           0x00BDE401
0x0270E1D0         1           0x00AD12D7
0x0270E1D0         2           0x00F8645F
0x0270E1D0         3           0x0157456A
0x0270E1D0         4           0x00408F5A
```

Scheduling

- Dynamic Affinity
  - ONLINE Schedulers
  - OFFLINE Schedulers
  - Hot Add CPU
- Worker limits
  - Max Worker Thread / ONLINE Schedulers
  - Ideal targets
- A Task
- A Worker
- Abort List
- I/O Completion Lists
- Idle Server
- Quantum Goals and Tracking (upyield vs SmartYield)

Scheduler Monitor

- Per node monitoring
- Checks for issues every 5 sec
  - 17883: Non-yielding task, single scheduler
  - 17887: I/O completion stall, single scheduler
  - 17884: No work progressing, all schedulers
  - 17888: 50% of same resource, all schedulers
- Uses installed callback routines
- CLR – Forced Yields – Garbage Collection
- Ring Buffer Entries

Visibility: Scheduler DMVs

- sys.dm_os_schedulers
- sys.dm_io_pending_io_requests
- sys.dm_os_workers
- sys.dm_os_tasks
- sys.dm_os_waiting_tasks

sys.dm_os_schedulers

- Similar to dbcc sqlperf(umsstats)
- Contains
  - Status
    - Visible
    - Online
    - Offline
    - DAC
  - Task counts
  - Worker Counts
  - Pending I/O Counts
  - Last Timer Activity (Join to: sys.dm_os_sys_info)
  - More…

sys.dm_io_pending_io_requests

- Outstanding I/O Requests
- Error Log Message for uncompleted > 15 seconds
- io_pending column = *HasOverlappedIoCompleted*
- Contains
  - Size
  - Offset
  - Time
  - More…

```
select fileInfo.*, pending.*
          from sys.dm_io_pending_io_requests as
   pending
   inner join (select * from
   sys.dm_io_virtual_file_stats(-1, -1)) as fileInfo
   on fileInfo.file_handle = pending.io_handle
```

sys.dm_os_workers

- All Workers
- One worker bound to one Task
- Contains
  - Worker Address
  - Task Association
  - Current State and Status
  - Performance Basics
  - Thread / Fiber Association
  - Scheduler Association
  - Quantum
  - More…

sys.dm_os_tasks

- All Tasks
- Contains
  - Task Address
  - Worker Association
  - Context Switches
  - Pending I/O Count
  - I/O Byte Counts
  - Session Id (Join to: sys.dm_exec_sessions)
  - More…

sys.dm_os_waiting_tasks

- Shows waiting tasks
- Preferable starting location to find waits
- Can grow large
- Contains
  - Core Task Information
  - Wait Type
  - Wait Duration
  - Session Id
  - More…

Buffer Pool

- Time of Last Access (TLA)
- Dynamic AWE
- I/O Affinity
- Hot Add Memory
- Copy On Write – Replica
- Checksuming
  - On Disk (default for all new databases)
  - In Memory (*-T831*)
  - - vs – Torn Page
  - Same as Exchange Server without automatic single bit correction
- Read Retry
- 823 (physical) and 824 (logical) failures
- Stalled I/O Detection (Data and Log)
- Stale Read Checks (*-T818*)
- Latch Enforcement (*-T815*) - Lightweight
- SQLIOSim.exe, replaces SQLIOStress.exe

Error 823 (Physical Failure)

- Occurs when: A *ReadFile*, *WriteFile*, *ReadFileScatter*, *WriteFileGather*, or *GetOverlappedResult* operation results in any *operating system error code*.

```
"The operating system returned error <<OS ERROR>> to
   SQL Server during a <<Read/Write>> at offset
   <<PHYSICAL OFFSET>> in file <<FILE NAME>>.
   Additional messages in the SQL Server error log and
   system event log may provide more detail. This is a
   severe system-level error condition that threatens
   database integrity and must be corrected
   immediately. Complete a full database consistency
   check (DBCC CHECKDB). This error can be caused by
   many factors; for more information, see SQL Server
   Books Online."
```

Error 824 (Logical Failure)

- **Logical Failures**
  - **Checksum**
  - **Torn Page**
  - **Short Transfer**
  - **Bad Page Id**
  - **Restore Pending (Enterprise Only)**
  - **Stale Read (-T818, also available in SQL Server 2000 SP4)**
  - **Page Audit (-T806, also available in SQL Server 2000 SP4)**

"SQL Server detected a logical consistency-based I/O error: <<ERROR TYPE DESCRIPTION>>. It occurred during a <<Read/Write>> of page <<PAGEID>> in database ID <<DBID>> at offset <<PHYSICAL OFFSET>> in file <<FILE NAME>>. Additional messages in the SQL Server error log or system event log may provide more detail. This is a severe error condition that threatens database integrity and must be corrected immediately. Complete a full database consistency check (DBCC CHECKDB). This error can be caused by many factors; for more information, see SQL Server Books Online."

sys.dm_os_buffer_descriptors

- Core information about HASHED buffers
- WARNING: List can be VERY LARGE
  - Avoid sorts
  - Select into doubles number of hashed buffers
- Immediate access required
- Enhancements likely

SQLIOSim.exe

- Part of HCL/HCT Testing
- No SQL Server components needed on machine
- Tested side by side with SQLIOStress.exe
- Sanity checks such as Checksum
- Simulates
  - DBCC Audit
  - Checkpoint
  - Lazy Writer
  - BCP / Select Into
  - Scanners
  - Random user activity
- Command Line or GUI Options
- 64 Bit Versions

Files and Configuration

Database Files

File: [ ] [...]

Size (MB): [ ] Max Size: [ ] Increment [ ]

[ ] Log File [ ] Shrinkable

File Name	Size (MB)	Max Size	Increment	Log File	Shrink
C:\SQLIOSim.mdx	50195	50195		FALSE	FALSE
C:\SQLIOSim.ldx	50195	50195		TRUE	FALSE

Add/Mod

Remove

System Level Configurations

CPUs: 2

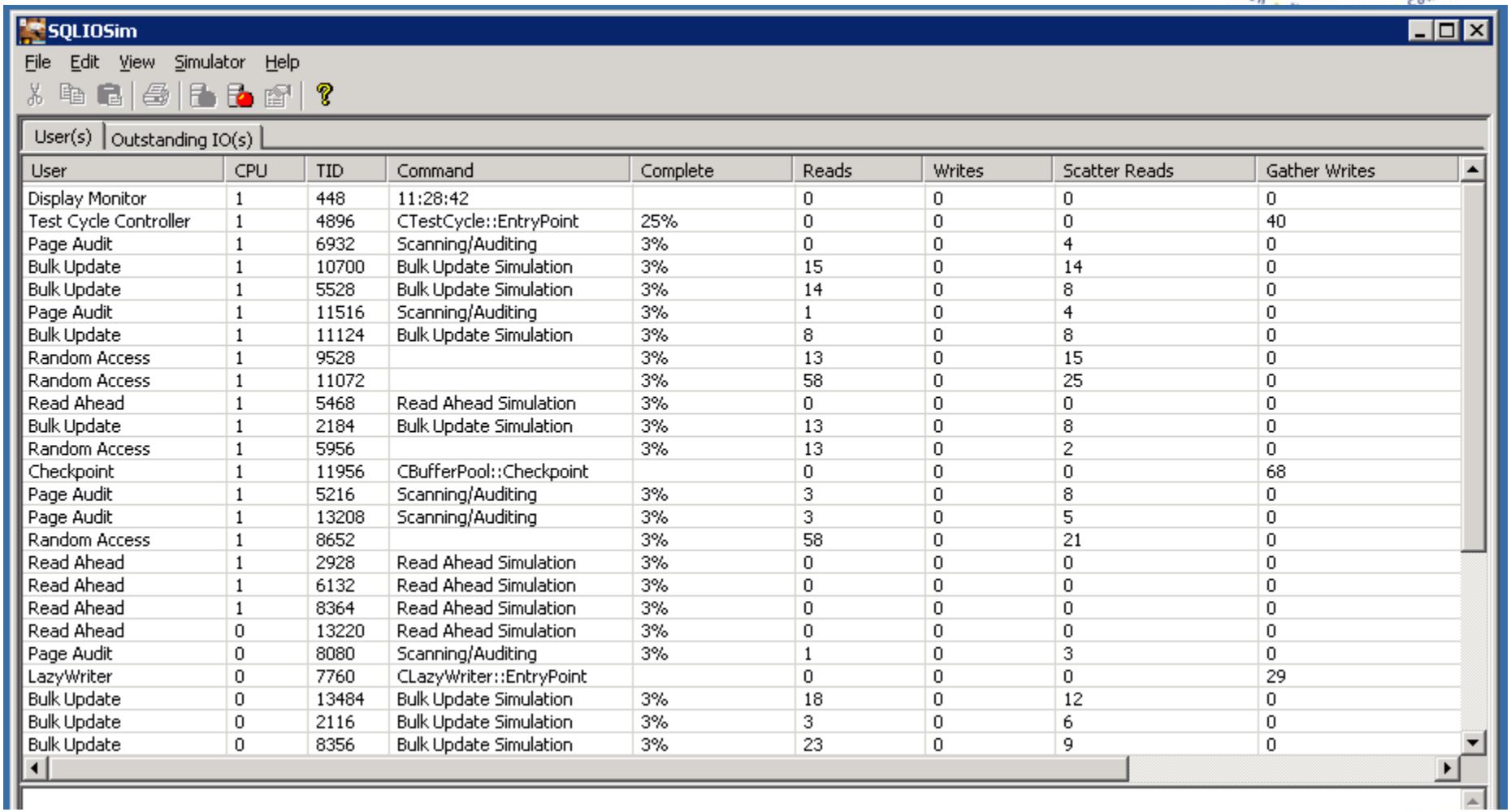
Affinity Mask: 0

IO Affinity Mask: 0

Max Memory (MB): 378

Error log: sqliosim.log.xml

OK Cancel



User	CPU	TID	Command	Complete	Reads	Writes	Scatter Reads	Gather Writes
Display Monitor	1	448	11:28:42		0	0	0	0
Test Cycle Controller	1	4896	CTestCycle::EntryPoint	25%	0	0	0	40
Page Audit	1	6932	Scanning/Auditing	3%	0	0	4	0
Bulk Update	1	10700	Bulk Update Simulation	3%	15	0	14	0
Bulk Update	1	5528	Bulk Update Simulation	3%	14	0	8	0
Page Audit	1	11516	Scanning/Auditing	3%	1	0	4	0
Bulk Update	1	11124	Bulk Update Simulation	3%	8	0	8	0
Random Access	1	9528		3%	13	0	15	0
Random Access	1	11072		3%	58	0	25	0
Read Ahead	1	5468	Read Ahead Simulation	3%	0	0	0	0
Bulk Update	1	2184	Bulk Update Simulation	3%	13	0	8	0
Random Access	1	5956		3%	13	0	2	0
Checkpoint	1	11956	CBufferPool::Checkpoint		0	0	0	68
Page Audit	1	5216	Scanning/Auditing	3%	3	0	8	0
Page Audit	1	13208	Scanning/Auditing	3%	3	0	5	0
Random Access	1	8652		3%	58	0	21	0
Read Ahead	1	2928	Read Ahead Simulation	3%	0	0	0	0
Read Ahead	1	6132	Read Ahead Simulation	3%	0	0	0	0
Read Ahead	1	8364	Read Ahead Simulation	3%	0	0	0	0
Read Ahead	0	13220	Read Ahead Simulation	3%	0	0	0	0
Page Audit	0	8080	Scanning/Auditing	3%	1	0	3	0
LazyWriter	0	7760	CLazyWriter::EntryPoint		0	0	0	29
Bulk Update	0	13484	Bulk Update Simulation	3%	18	0	12	0
Bulk Update	0	2116	Bulk Update Simulation	3%	3	0	6	0
Bulk Update	0	8356	Bulk Update Simulation	3%	23	0	9	0

Microsoft SQL Server© Simulator Stress Test Results

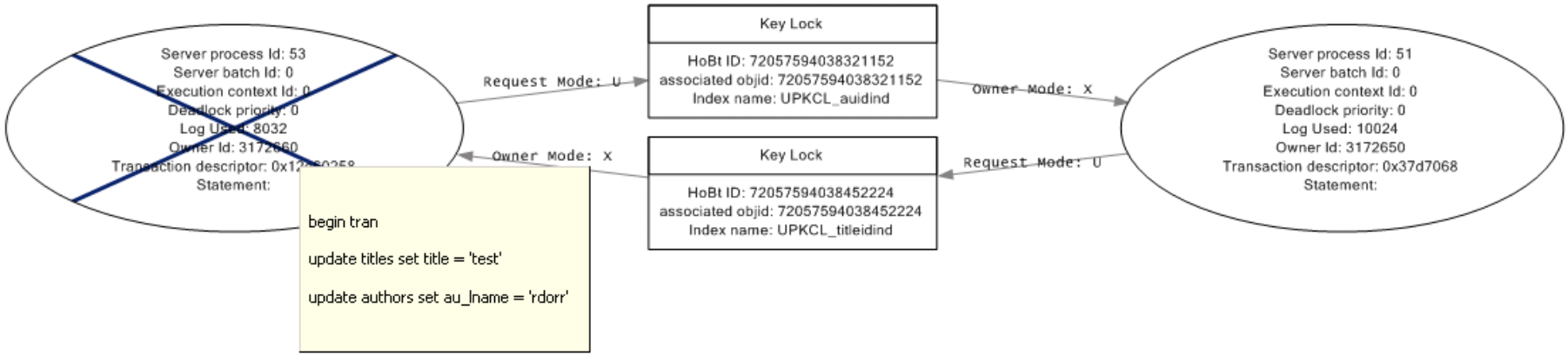
Date	Time	Tid	User	Description
08/26/05	11:33:44	11444	System	Starting Microsoft SQL Server(c) Simulator Stress Test Version 1.00.000
08/26/05	11:33:44	11444	System	Logical CPUs: 2
08/26/05	11:33:44	11444	System	Affinity: 0
08/26/05	11:33:44	11444	System	IO Affinity: 0
08/26/05	11:33:44	14148	CreateFileStream	Creating file c:\temp\test.mdf\sqliosim.ldx
08/26/05	11:33:44	14148	CreateFileStream	Error: 0x80070003 Error Text: The system cannot find the path specified. Description: API: CreateFile FILE: c:\temp\test.mdf\sqliosim.ldx
08/26/05	11:33:44	6520	CreateDB	Destroying file c:\temp\test.mdf\sqliosim.mdx
08/26/05	11:33:44	6520	CreateDB	Destroying file c:\temp\test.mdf\sqliosim.ldx
08/26/05	11:33:44	2924	Display Monitor	Cleaning up buffer pool
08/26/05	11:33:44	2924	Display Monitor	Buffer Pool: validated buffers 0, pages 0, discarded buffers 0
08/26/05	11:33:44	2924	Display Monitor	Closing file c:\temp\test.mdf\sqliosim.mdx
08/26/05	11:33:44	2924	Display Monitor	********** Final Summary for file c:\temp\test.mdf\sqliosim.mdx **********
08/26/05	11:33:44	2924	Display Monitor	File Attributes: Compression = No, Encryption = No, Sparse = No
08/26/05	11:33:44	2924	Display Monitor	Target IO Duration (ms) = 25, Running Average IO Duration (ms) = 25, Number of times IO throttled = 0, IO request blocks = 1
08/26/05	11:33:44	2924	Display Monitor	Reads = 0, Scatter Reads = 0, Writes = 0, Gather Writes = 0, Total IO Time (ms) = 0
08/26/05	11:33:44	2924	Display Monitor	Error: 0x00000000 Error Text: Description: Drive does not appear to be of type Fixed Media
08/26/05	11:33:44	2924	Display Monitor	DRIVE LEVEL: Sector size = 0, Cylinders = 0, Media type = 0, Sectors per track = 0, Tracks per Cylinders = 0
08/26/05	11:33:44	2924	Display Monitor	DRIVE LEVEL: Read cache enabled = No, Write cache enabled = No
08/26/05	11:33:44	2924	Display Monitor	DRIVE LEVEL: Read count = 325736, Read time = 19184448, Write count = 2503060, Write time = 6027157, Idle time = 1363246694, Bytes read = 492


```
C:\Temp\SQLIOSim>sqliosim -log c:\temp\delme.xml
User                    Information                         Complete
----                    -----------                         --------
Main User               Refreshed 16 times
Read Ahead              0:384, 0 pending                      4%
Read Ahead              0:0, 0 pending                        4%
Read Ahead              0:512, 0 pending                      4%
Read Ahead              0:256, 0 pending                      4%
Page Audit              0:600                                 4%
Page Audit              0:584                                 4%
Bulk Update             0:525, Reading page(s)                4%
Bulk Update             Sleeping 10 ms                        4%
Bulk Update             0:2, Reading page(s)                  4%
Bulk Update             0:476, Reading page(s)                4%
Random Access           0:446, Updating page(s)               4%
Random Access           0:290, Updating page(s)               4%
Page Audit              0:592                                 4%
Random Access           0:385, Updating page(s)               4%
LogWriter               Sleeping, 15775 processed
Page Audit              0:576                                 4%
LazyWriter              Sleeping, 589 modified
Random Access           0:447, Updating page(s)               4%
Random Access           0:485, Reading page(s)                4%
Read Ahead              0:256, 0 pending                      4%
Read Ahead              0:128, 0 pending                      4%
Read Ahead              0:384, 0 pending                      4%
Read Ahead              0:128, 0 pending                      4%
Page Audit              0:544                                 4%
Page Audit              0:552                                 4%
Bulk Update             0:488, Reading page(s)                4%
Bulk Update             Sleeping 10 ms                        4%
Bulk Update             Sleeping 10 ms                        4%
Bulk Update             0:533, Reading page(s)                4%
Random Access           0:3, Reading page(s)                  4%
Random Access           0:441, Updating page(s)               4%
Page Audit              0:560                                 4%
Random Access           0:313, Updating page(s)               4%
Checkpoint              Sleeping
Test Cycle Controller   Full Test Run                        25%
Page Audit              0:568                                 4%
Display Monitor         Sleeping
```

Lock Monitor

- Deadlock Detection Extended
  - Hosted Objects
  - CLR Objects
- Blocked Process Threshold
  - sp_configure option
  - Broker Event
  - Trace Event
  - XML Data
  - Detection less than 5 seconds is irregular
- Deadlock Information
  - XML Output
  - GUI Support

Visual Deadlock



Latch Statistics

- Like dbcc sqlperf(waitstats) for latches
- Companion to sys.dm_os_wait_status
- Contains
  - Latch Class
  - Request Count
  - Wait Time
  - Max Wait Time

Hosting

- Core Interfaces
  - Threading
  - Synchronization Objects
  - Memory Allocations
  - More…
- Detour Testing
- SOSBOOT.dll – DLLMain

Mini-Dumps

- Indirect Memory
- All Threads
  - 17883 – Stuck scheduler
  - 17884 – All schedulers stuck
  - 17887 – Stuck I/O completion routine
  - 17888 – 50% of workers all have same wait
  - CLR Core Exceptions (Fatal)
  - Unexpected termination calls
  - More…
- Enhanced mini-dumps
  - Ring buffers by default
  - More…
- Filtered Dumps
- Watson
- Cluster Failover Capabilities

Memory

- New caching facilities
- Supports For Hosted Memory Consumers
- LRU Based on Allocation Cost
- Resource Monitor
- Increased Visibility
- http://blogs.msdn.com/slavao

Old Style: dbcc memorystatus

- Global Details

```
Memory Manager                  KB
------------------------------- --------------------
VM Reserved                     1062016
VM Committed                    28892
AWE Allocated                   0
Reserved Memory                 1024
Reserved Memory In Use          0
```

- Per Node Details

```
Memory node Id = 0              KB
------------------------------- --------------------
VM Reserved                     1057920
VM Committed                    24948
AWE Allocated                   0
MultiPage Allocator             7696
SinglePage Allocator            6608
```

- Per Clerk Details

sys.dm_os_memory_clerks

- Details about each clerk
- ~131 Clerks rows
- Clerks = Accountants
- Contains
  - Type
  - AWE Information
  - Node Id
  - Single Page Allocations
  - Multi-Page Allocations
  - More…

```
type                         memory_node_id single_pages_kb      multi_pages_kb
---------------------------- -------------- -------------------- ----------------
MEMORYCLERK_SQLOPTIMIZER     0              248                  72
MEMORYCLERK_SQLUTILITIES     0              72                   0
MEMORYCLERK_SQLSTORENG       0              1360                 88
```

Drilling Into Memory

- WARNING: Memory DMVs can become very large
- Performance sensitive columns and views are only populated when trace flag *3654* is enabled
- Clerks (sys.dm_os_memory_clerks)
  - Objects (sys.dm_os_memory_objects)
    - Allocations (sys.dm_os_memory_allocations)

Clerks to Objects

- Join column: page_allocator_address

```
select …
 from sys.dm_os_memory_clerks c
 join sys.dm_os_memory_objects m on
   c.page_allocator_address = m.page_allocator_address
```

- Example after select * from sys.dm_os_schedulers executed

```
memory_object_address type         ClerkType        ClerkName    name

-------------------- ------------ --------------- ------------ ----------------------
0x11BF02D0           MEMOBJ_PARSE CACHESTORE_PHDR Bound Trees  Parse_dm_os_schedulers
```

Objects To Allocations

- sys.dm_os_memory_allocations requires trace flag 3654
- Large Data Sets
- Join column: memory_object_address

```
select m.name, a.* from sys.dm_os_memory_objects m
   join sys.dm_os_memory_allocations a
   on m.memory_object_address = a.memory_object_address
   join sys.dm_os_memory_clerks c on
   c.page_allocator_address = m.page_allocator_address
```

- Example Output of a sys.dm_os_schedulers execution

Column	Value
name	Parse_dm_os_schedule
creation_time	2005-09-02 11:12:27.733
allocator_stack_address	0x0F78A580
source_file	t:\yukon\sql\ntdbms\msql\norm\algrel.cpp
line_num	1234

Allocations to Stack

- Requires trace flag *3654*
- Large Data Sets
- Join columns: stack_address, allocator_stack_address

```
select * from sys.dm_os_stacks s
  join sys.dm_os_memory_allocations a on
  a.allocator_stack_address = s.stack_address
  join  sys.dm_os_memory_objects m on
  m.memory_object_address =
  a.memory_object_address
order by s.frame_index
```